

NASA TP 3665



Independent Verification and Validation of Complex User Interfaces: A Human Factors Approach

Mihriban Whitmore
Andrea Berman
and
Cynthia Chmielewski

December 1996

1N-61
021636

NASA TP 3665

Independent Verification and Validation of Complex User Interfaces: A Human Factors Approach

Mihriban Whitmore
Andrea Berman
and
Cynthia Chmielewski
*Lockheed Martin Space Mission Systems & Services Company
Houston, Texas*

December 1996

National Aeronautics
and Space Administration

Lyndon B. Johnson Space Center
Houston, Texas

ACKNOWLEDGMENTS

This research was supported by Contract Number NAS9-18800 from the National Aeronautics and Space Administration and conducted at the Lyndon B. Johnson Space Center in Houston, Texas. The authors wish to thank Rebeka Burnett of the Flight Crew Support Division for her support in preparing the interfaces for evaluation and in training the experimenters on the Sun Operating System. Jurine Adolf and Carlos Sampaio of Lockheed Martin Space Mission Systems and Services Company (LMSMSSC), Tina Holden of Syntropy Corporation, and Justin Miller, currently of Apple Corporation, are thanked for their participation in either one or both phases of this experiment. Jennifer Novak and Jacquie Minton of LMSMSSC are thanked for their data collection support. Thanks are also extended to Manny Diaz and Jan Panero for their efforts in the early conception of the project.

This publication is available from the NASA Center for AeroSpace Information,
800 Elkridge Landing Road, Linthicum Heights, MD 21090-2934, (301) 621-0390.

CONTENTS

	<u>PAGE</u>
ACKNOWLEDGMENTS.....	ii
CONTENTS	iii
ACRONYMS	v
ABSTRACT.....	vi
1.0 INTRODUCTION	1
1.1 Background.....	1
1.2 IV&V Methodology	2
1.3 CHIMES	3
1.4 KRI/AG.....	4
1.5 University of Maryland Tool	5
2.0 OBJECTIVES.....	7
3.0 PHASE I.....	7
3.1 Methods	8
3.1.1 Participants	8
3.1.2 Apparatus and Materials.....	9
3.1.3 Procedure	9
3.1.3.1 CHIMES Compliance Check.....	9
3.1.3.2 Heuristic Evaluation	10
3.2 Results.....	10
3.2.1 Guideline Compliance Reporting	10
3.2.2 Completion Time Results.....	11
3.3 Discussion.....	11
4.0 PHASE II.....	12
4.1 Methods	12
4.1.1 Participants	12
4.1.2 Apparatus and Materials.....	12
4.1.3 Procedure	12
4.2 Results.....	13
4.2.1 Fields/Buttons	13

CONTENTS

(continued)

	<u>PAGE</u>
4.2.2 Labeling	14
4.2.3 Feedback	14
4.2.4 Procedures	15
4.3 Discussion.....	16
5.0 CONCLUSIONS.....	17
6.0 RECOMMENDATIONS FOR FUTURE WORK.....	17
REFERENCES.....	18
APPENDIX A.....	A-1
APPENDIX B.....	B-1
APPENDIX C.	C-1
APPENDIX D	D-1
APPENDIX E.....	E-1
APPENDIX F.....	F-1
APPENDIX G	G-1

Figures

1	A display of IV&V use at various stages of the design process	3
2	Percentage of selected problem categories reported across all four displays by CHIMES and by HCI experts.....	10
3	Summary of results from the Phase II questionnaire.....	16

Tables

1	Comparison of Existing IV&V Tools	6
2	HCI Expert Backgrounds for Participants in Phases I and II.....	8

ACRONYMS

CHIMES	Computer Human Interaction Models
GUI	graphical user interface
ISS	International Space Station
IV&V	independent verification and validation
HCI	human-computer interface
KRI/AG	Knowledge-Based Review of User Interfaces
PCS	Portable Computer System
TAE+	Transportable Applications Environment
UIMS	user interface management system
UM	University of Maryland
UTAF	Usability Testing and Analysis Facility

ABSTRACT

The Usability Testing and Analysis Facility (UTAF) at the NASA Johnson Space Center has identified and evaluated a potential automated software interface inspection tool capable of assessing the degree to which space-related critical and high-risk software system user interfaces meet objective human factors standards across each NASA program and project. Testing consisted of two distinct phases. Phase I compared analysis times and similarity of results for the automated tool and for human-computer interface (HCI) experts. In Phase II, HCI experts critiqued the prototype tool's user interface. Based on this evaluation, it appears that a more fully developed version of the tool will be a promising complement to a human factors-oriented independent verification and validation (IV&V) process.

Independent Verification and Validation of Complex User Interfaces: A Human Factors Approach

1.0 INTRODUCTION

1.1 Background

The user interface is a critical part of any computer system and merits careful evaluation before it is released to users. The independent verification and validation (IV&V) methodology represents a series of activities which strive to improve the quality and the reliability of user interfaces and to ensure that the delivered interfaces satisfy the users' operational needs. The goal of IV&V is to carefully ensure quality software throughout a complete system.

Lewis (1992) indicates that IV&V should concentrate on how displays look, how they are controlled, and the quality of engineering that goes into them. One way to evaluate interfaces is to check them for compliance with standards and guidelines. However, this is not a simple or a fail-safe process.

First, it is often difficult for designers to follow guidelines in the initial design phase. De Souza & Bevan (1990) found that designers had difficulties with or made errors in using an average of 66% of the guidelines they were given. Nearly every guideline (91%) caused some difficulty for at least one designer in the study. Tetzlaff & Schwartz (1991) noticed similar difficulties. Both studies found that designers were somewhat better at using the guidelines in prototypes than they were at understanding the guidelines. In other words, designers displayed misunderstanding of the guidelines through protocols and debriefing interviews but showed moderate conformance nonetheless.

Problems are also found with assessing conformance with guidelines. Thovtrup & Nielsen (1991) found that experienced designers with an interest in human factors found an average of only 4 out of 12 violations of guidelines in four screen dumps. Another study found that evaluators tended to use their personal assessments of usability over actual judgment of conformity, or compensated for the errors of designers through their own understanding, thus not reporting violations of the guidelines (Tetzlaff & Schwartz, 1991).

Furthermore, procedures for ensuring that human factors guidelines are met in the design of human-computer interfaces have been mostly accomplished on a case-by-case basis. Many human factors concepts (i.e. how displays are controlled) are subjective or complicated. This issue may lead to inconsistencies in the application of human factors guidelines and, in some cases, failure to incorporate these guidelines.

1.2 IV&V Methodology

Researchers suggest a number of methods for improving the usability of guidelines and standards. These include clarifying the conditions under which a guideline should be applied, including explanations of terminology, providing procedures for determining thresholds for items such as "frequency of use," and including information on the intent of the scope of the guideline (de Souza & Bevan, 1990).

Other suggestions include adding examples to general guidelines, making standards accessible on-line so that they may be searched, and providing priorities and reliability ratings for each guideline (Mosier & Smith, 1986). Tetzlaff & Schwartz (1991) indicate that important guidelines should have graphical examples. However, added examples, explanations, priorities, and severity ratings do not help if, as some studies have found, designers do not read the guidelines or standards (Lowgren & Lauren, 1993; Tetzlaff & Schwartz, 1991), and evaluators allow their subjective experiences to prevail over actual compliance or noncompliance (Tetzlaff & Schwartz, 1991; Thovtrup & Nielsen, 1991).

Another way to make the application of guidelines and standards more accessible and effective is to integrate automaticity into the process—an impartial judge to call guideline violations to the attention of designers and evaluators. Figure 1 depicts how such a tool can be utilized at various stages during the design process (adapted from Baecker, Grudin, Buxton, & Greenberg, 1995). The first phase, conceptual design, includes the human-only processes of information collection and participative design. However, in phase 2 when the actual interface design begins and a prototype is implemented, the addition of an automated IV&V tool would be very advantageous. Using an automated tool to perform the task of checking human factors guidelines would be both cost- and time-efficient, thereby maximizing productivity and quality. An IV&V tool can also be useful during the enhancement and evolution of the system in Phase 3. Such a tool could effortlessly and efficiently check that new features are compatible with the original system. Sears (1994) suggests that two types of automated metrics be used to evaluate interfaces. The first would be task-sensitive metrics which ensure that interfaces are appropriate for the users' tasks. He proposes an algorithm, *Layout Appropriateness*, which calculates the efficiency of the organization of the objects based on size, distance, and frequency of use. The second set of metrics would be for task-independent evaluations, focusing on the general appearance of the interfaces. Sears (1994) suggests these metrics be based on the work of Tullis (1983), who has developed algorithms to measure properties such as overall screen density, local density, grouping, and layout complexity. Tullis' work is based primarily on alphanumeric displays; however, there are several tools that have been developed for graphical displays which implement automated checking against human factors guidelines and standards.

The Uses of IV&V in the Design Process

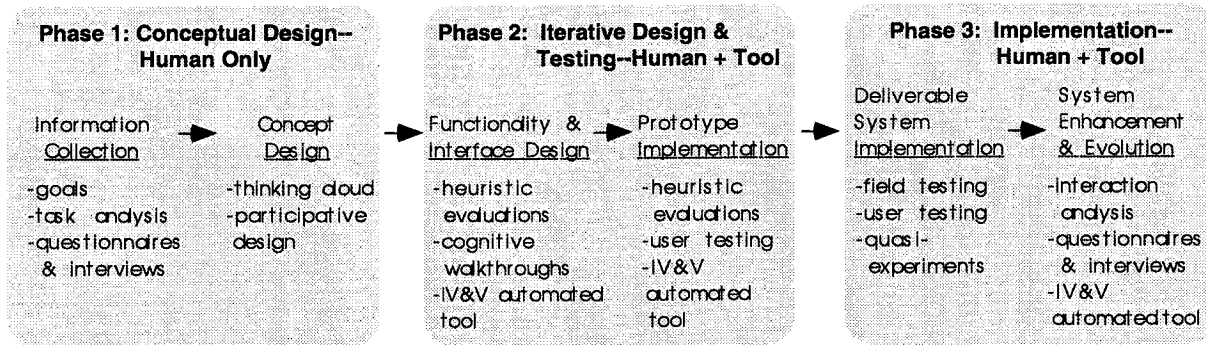


Figure 1: A display of IV&V use at various stages of the design process.

The design of an expert system can be complicated. Using experts to evaluate displays throughout each step of the design process can be costly and time-consuming. Therefore, it would be advantageous to include the use of an automated IV&V tool to complement designers and expedite the evaluation process. The Usability Testing and Analysis Facility (UTAF) at the National Aeronautics and Space Administration (NASA) Lyndon B. Johnson Space Center (JSC) initially reviewed the features of three tools developed for graphical displays—CHIMES (Jiang, Murphy, & Carter, 1994), KRI/AG (Lowgren & Lauren, 1993; Lowgren & Nordquist, 1992), and the University of Maryland Tool (Shneiderman, Chimera, Jog, Stimart, & White 1995; Mahajan & Shneiderman, 1995).

1.3 CHIMES

Computer-Human Interaction Models (CHIMES) was developed at the NASA Goddard Space Flight Center. The CHIMES tool was created to demonstrate the feasibility of automating user interface evaluations from the perspective of human factors guidelines and heuristics. The primary goal of this tool is to check objective interface characteristics such as button size, labeling and location, fonts, point sizes, and the use of colors for conformance to standards. The human-computer interface (HCI) expert is then spared those time-consuming, monotonous tasks and can concentrate on more cerebral interface issues, such as functionality and interaction behaviors.

CHIMES is able to switch between two modes, depending on whether an alphanumeric interface or a graphical user interface (GUI) is being checked. In the demand-modeling mode used for alphanumeric interfaces, CHIMES estimates the demands that the interface will place on an experienced operator's multiple cognitive resources. In the guidelines-based GUI mode, CHIMES checks for compliance with human factors guidelines and toolkit style requirements. Unfortunately, the demand-modeling mode

was unavailable for evaluation (Harris, 1996). Therefore, the remainder of this evaluation concentrates on the GUI mode.

The guidelines-based approach to GUI evaluation is conceptually similar to providing a spell checker or a grammar checker for a text document. Having designed a GUI component, the designer invokes CHIMES to check the design for compliance with various sets of GUI guidelines. In addition, CHIMES allows users to customize its rules and to focus on particular interface components (e.g. button sizes). Knowledge-based design advice from information contained in the human factors literature is provided by CHIMES if any non-compliance is detected. The CHIMES user manual includes a detailed description of the references for the human factors guidelines (Jiang, Murphy, & Carter, 1994). CHIMES provides a short list of all advice produced and then the user can choose any item from the list and display more detailed contents. Advice is also provided in context when it pertains to a particular object in the interface (a pointer is displayed next to the object in question). Next, the designer can make changes in the GUI component or, in some cases, can request CHIMES to make the appropriate changes. After making changes, the designer can re-invoke CHIMES, and this process can continue until the designer is satisfied that the design complies with known guidelines. The capabilities and limitations of this tool are summarized in Table 1.

1.4 KRI/AG

Knowledge-Based Review of User Interfaces (KRI/AG) is an expert system which evaluates GUIs (Lowgren & Lauren, 1993; Lowgren & Nordquist, 1990; 1992). KRI/AG runs on Motif interfaces developed under the TelUSE user interface management system (UIMS). This system uses knowledge from standards and Motif style guides to uncover problems with presentation and display syntax. The system is fairly interactive; it steps through the interface, generating conclusions and messages. If the same flaw is found throughout an interface, the comments produced are aggregated into one. The evaluator may: 1) select domains to be examined further, 2) ask for reasons for the comments, 3) ask for suggestions for improvement, and 4) ask for direct quotes and references from guidelines that are violated. Users of KRI/AG have complained that many of the comments are trivial or fail to take user and task characteristics into account. One difficulty with implementing such an expert system is that many guidelines are general or simply difficult for a computer to interpret (e.g., "related items should be grouped from general to specific"). Such guidelines tend to cause the production of comments considered trivial or irrelevant to the evaluator.

Through the use of the KRI/AG, Lowgren & Nordquist (1992) and Lowgren & Lauren (1993) have a number of suggestions for the improvement of automated interface-checking tools. One is to expand the tool for task-dependent evaluation by integrating runtime usage logs into the tool. Data could be collected from users as they perform tasks, and KRI/AG could make recommendations based on selection frequencies,

common sequences, errors, help requests, and so on. Analysis could be done with an algorithm such as maximal repeating pattern analysis described by Siochi & Hix (1991), which determines patterns of system use by recording what users do. Other suggestions by Lowgren & Lauren (1993) are preferences from designers for increasing the acceptability of automated evaluation tools. Designers in this study indicated that they would not use a tool which persisted in giving many comments with which they disagreed. Since strict evaluation is likely to give results with which designers will disagree (similar to grammar checkers that complain about the length of sentences over a set number of words), this concern is not unreasonable. Lowgren & Lauren (1993) report that two of the most critical concerns are that the tool must provide information on the level of severity of a violation and that it must be configurable by its users. Possible configurations include setting the tool to skip certain types of comments or particular guideline violations. For example, if evaluators have agreed that a certain piece of software will be exempted from a particular guideline, messages about violations of that guideline will not appear repeatedly. Users may also want to be able to choose among different levels of stringency, so that they might only see messages about very severe violations on a first pass. Stringency may then be increased for more developed interfaces. The capabilities and limitations of this tool are summarized in Table 1.

1.5 University of Maryland Tool

Another tool designed for interface checking is under development at the University of Maryland (UM). This tool currently evaluates only dialog boxes, not the entire interface, and it is not yet interactive or automatic. The tool provides a way to compare all the dialogs in an interface and check them for consistency and conformance to guidelines. The tool must now be used non-interactively. That is, information is given on summary printouts. The printouts must be scanned manually for patterns and anomalies; knowledge of guidelines is not yet integrated into the tool. Automation will be added later (Mahajan & Shneiderman, 1995; Shneiderman, Chimera, Jog, Stimart, & White, 1995). The tool operates on interfaces created in Visual Basic, but interpreters can be written so that any interface may be translated to the canonical form read by the tool.

The UM tool consists of six parts. The dialog box typeface and color schemes create a table with one row for each dialog in the interface. Three columns list the name of the dialog box, the typefaces used (distinct typefaces include variations in font, size, and style), and the colors used. Each distinct typeface and color is keyed to a number so that the table output is compact and the number of typefaces and colors used within each box as well as throughout the interface is clear.

A second part of the tool is the interface concordance, which examines all text used in the dialog boxes. The output consists of any words that have "variant capitalization" ("quit" vs. "Quit" vs. "QUIT"), variant pluralization ("quit" vs. "quits"), and variant

punctuation (“quit” vs. “quit:”). Each form of the words in question is printed out, followed by the names of the dialogs in which they occur. The button concordance performs a similar analysis, however it focuses only on labels found on buttons, since consistency here may be especially crucial.

The button layout table examines groups of buttons. The grouping must be defined ahead of time and includes all labels that might be used together, including synonyms. A set of terms might be “Add, Remove, Delete, Copy, Clear, Cancel, Close, Exit.” The tool finds all of the actual groups of buttons in the dialog boxes (one group might be “Add, Remove, Exit”) and lists the sizes and relative positions of the buttons in the group. This portion of the tool, therefore, informs the user of which buttons have been used together (given which buttons had been predicted to be used together) and whether there is consistency in sizes and spacing of buttons within and across groups.

The interface speller lists words used in the dialog box which are not found in the dictionary. This capability is not only useful in finding typographical errors, but also in detecting potentially confusing abbreviations.

The final section of this tool is the basket browser. Terminology baskets are formed as a kind of thesaurus of interface terms. For example, one basket may be “Search Retrieve Query Select.” The terminology baskets ignore variations in capitalization and punctuation and list cases where the different words are used for the same task in different dialogs, thus further promoting consistency across the interface. Table 1 summarizes the capabilities and limitations of this tool.

Table 1: Comparison of Existing IV&V Tools

Tool	Capabilities	Limitations
CHIMES	<ul style="list-style-type: none"> • Fairly interactive • Assesses consistency of objective interface characteristics (i.e., fonts, buttons) 	<ul style="list-style-type: none"> • Only runs on a Sun workstation with SunOS, Motif, and TAE+ User Interface Management System
KRI/AG	<ul style="list-style-type: none"> • Fairly interactive • Multiple comments on the same topic are compiled together 	<ul style="list-style-type: none"> • Only runs with Motif • Attempts to verify general guidelines
UM Tool	<ul style="list-style-type: none"> • Assesses consistency of objective interface characteristics (i.e., fonts, buttons) • Interfaces can be translated into a format to be read by the tool 	<ul style="list-style-type: none"> • Not yet interactive and automatic • Currently checks only dialog boxes

The UTAF has been tasked with identifying and evaluating a potential IV&V software tool. This tool should be capable of assessing the degree to which space-related critical and high-risk software system user interfaces meet human factors standards across each NASA program and project, such as the International Space Station (ISS). After reviewing literature on each of the tools and their capabilities, the CHIMES tool was found to be most acceptable for evaluation. Since the CHIMES prototype was interactive and able to check the consistency of objective interface characteristics across a series of displays, it best met UTAF requirements. In addition, it was easily accessible as NASA-developed software.

2.0 OBJECTIVES

The objectives were to evaluate the performance, usability, and the interface design of the CHIMES tool. Two separate evaluations were conducted to achieve these goals:

- Phase I: The ability of CHIMES to evaluate an interface as compared to an HCI expert was assessed. In addition, the evaluation sought to reveal certain features required in an ideal IV&V tool.
- Phase II: HCI experts examined the user interface and potential applications of the CHIMES tool.

3.0 PHASE I

Initial subjective investigations showed that CHIMES was faster than an HCI expert at checking objective interface characteristics, such as number of typefaces per display, type size, type style, line thickness, line style, color usage, and consistency of each across multiple interfaces. These initial results led to Phase I of the CHIMES usability testing process.

The primary goals of the first phase were to refine the ideal IV&V tool concept and methodology and to define the functional requirements of the tool. The specific usability questions asked were:

- Does CHIMES perform as well as HCI experts when evaluating highly complex user-interfaces?
- How long does it take for CHIMES and for HCI experts to complete a human factors compliance and consistency evaluation?

3.1 Methods

3.1.1 Participants

Four subjects participated in this phase of the evaluation. Two were employees of a contracting organization with JSC. One was a previous employee of a contracting organization with JSC, and the fourth had been a Rice University intern in the UTAF. Table 2 includes a brief description of both Phase I and Phase II participants' HCI knowledge.

Table 2: HCI Expert Backgrounds for Participants in Phase I and II

Experts*:	1	2	3	4	5
Number of operating systems known	6	2	4	3	6
Usability activities					
• Participated as a subject	5 Xs	3 Xs	>10 Xs	6 Xs	3 Xs
• Conducted software usability testing	8 Xs	1 Xs	>10 Xs	3 Xs	3 Xs
• Designed software usability experiments	6 Xs	N/A	>10 Xs	16 Xs	3 Xs
• Designed software interfaces	6.5 yrs	N/A	>10 Xs	20 Xs	N/A
• Checked interfaces for HCI compliance	5 yrs	4 Xs	3-4 Xs	12 Xs	N/A
• Performed display reviews	4 yrs	4 Xs	2-3 Xs	12 Xs	N/A
• Performed rapid prototyping	7 yrs	2 Xs	>10 Xs	12 Xs	N/A
Familiarity with ISS displays (Yes/Y or No/N)					
• Conducted evaluation of all subsystems	Y	N/A	N/A	Y	N/A
• Conducted evaluation of one subsystem	Y	Y	N/A	N/A	N/A
• Attended crew reviews	Y	Y	N/A	Y	N/A
• Attended display demonstrations	Y	Y	N/A	Y	N/A
• Participated in display & control meetings	Y	Y	N/A	Y	N/A
• Other ISS display efforts	Y	N/A	N/A	N/A	N/A
HCI Background					
• Education in HCI	5 yrs	3 yrs	5 yrs	7 yrs	1.5 yrs
• Technical training	4 yrs	3 yrs	4 yrs	N/A	N/A
• Education in human factors	6 yrs	3 yrs	5 yrs	7 yrs	2.5 yrs
• Work experience	6.5 yrs	8 yrs	4 yrs	9 yrs	1.5 yrs
Familiarity with HCI documentation (Familiar/F or Extensive Use/E)					
• Macintosh Human Interface Guidelines	E	F	E	F	F
• Object-oriented Interface Design	E	F	F	N/A	F
• Open Look: Graphical User Interface Application Style Guidelines	F	N/A	F	N/A	N/A
• Other documentation	E	N/A	N/A	F	N/A
Participation in Phases I and II	I,II	I,II	I	I	II

*experience is given in years (yrs) or times involved with the activity (Xs)

All four participants had HCI graduate education and a minimum of two years of experience in usability testing and in HCI reviews. Three of the experts were involved in interface design, and three of the experts were familiar with ISS displays. All had computer experience.

3.1.2 Apparatus and Materials

CHIMES runs under the Transportable Applications Environment (TAE+) UIMS. It requires a SUN workstation with SunOS and Motif and can only evaluate interfaces developed in TAE+.

CHIMES consists of three primary screens: "Evaluation Control" to identify evaluation categories, to select the resource file, and to initiate the evaluation process; "Advice Index" to view the list of problems; and "Problems and Advice" to view a more detailed description of a particular problem and to receive advice. In addition, for certain design problems, the user may access a "modifier" screen to actually modify the design during an evaluation. For more detailed information on CHIMES refer to the user manual (Jiang, Murphy, & Carter, 1994).

Four fairly mature ISS Portable Computer System (PCS) user interface designs were evaluated: 1) top-level PCS display, 2) lab fire display, 3) mobile transporter home display, and 4) habitation module overview display (see Appendix A). These interfaces were selected from the PCS Team website for their highly complex nature. The displays were re-created using TAE+ and were merely screen images with no interactivity.

3.1.3 Procedure

The interface evaluation consisted of two parts: the CHIMES compliance and consistency check and the HCI expert review of the displays. Performance measures used for comparison of CHIMES to the HCI experts were the percentage of possible HCI-related guideline violations reported and the total task completion time. Mean expert completion times were compared to the total CHIMES time, and the results of each expert's analysis were cross-checked with CHIMES' results.

3.1.3.1 CHIMES Compliance Check

For the CHIMES compliance check, the four displays were evaluated separately to determine the HCI guidelines violations. They then were evaluated simultaneously for the consistency check across interfaces. Evaluation control information was entered (e.g., source file, advice categories), and then CHIMES performed the evaluation and generated its "advice index." Each problem-advice pair was recorded, and the completion time for the evaluation was noted.

3.1.3.2 Heuristic Evaluation

The HCI experts individually performed the same heuristic evaluation on each interface as well as a consistency check across interfaces. They followed a formal protocol to ensure that each expert checked the same objective characteristics. The experts were provided with a list of general design items of concern to interface designers and were asked to critique the interfaces in terms of those design items (e.g., font use, button design, line use, and color use). They were provided with tables in which to enter their free-form display comments (see Appendix B) and also rated each display design item on a five-point rating scale anchored as 1=Redesign from Scratch and 5=Design is Completely Acceptable. After the evaluation, the experts rated the procedure on a structured questionnaire with a five-point rating scale anchored as 1=Strongly Disagree and 5=Strongly Agree (see Appendix C).

3.2 Results

The results for both performance measures are presented below in separate sections. The HCI experts' and CHIMES' detailed critiques of the ISS interfaces are reported in Appendix D.

3.2.1 Guideline Compliance Reporting

Figure 2 shows the results for the guideline compliance portion of the Phase I testing. For each of the interface problem categories, the majority of items were reported by both CHIMES and the experts. However, there were also many items that only were detected by one or by the other.

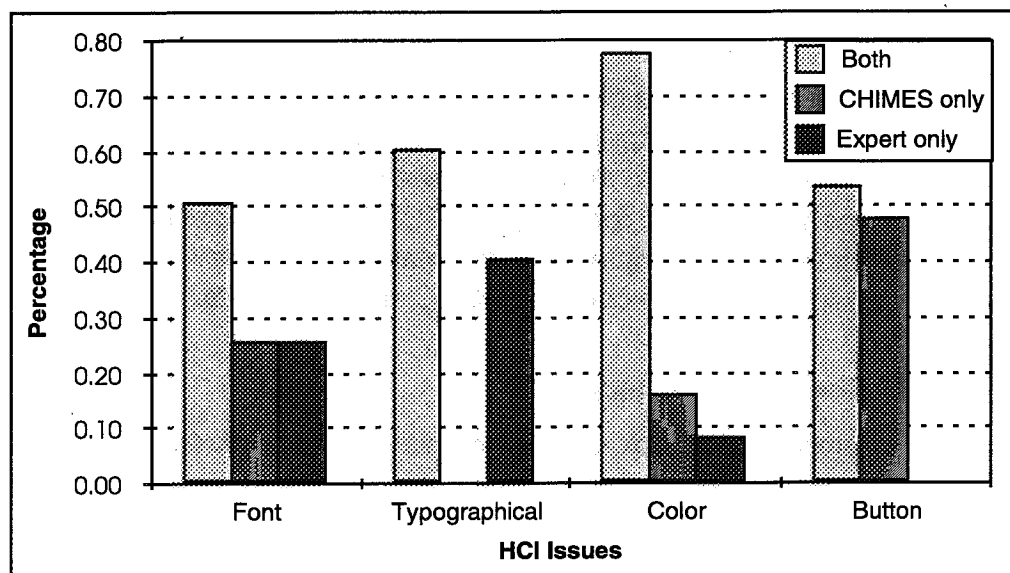


Figure 2: Percentage of selected problem categories reported across all four displays by CHIMES and by HCI experts.

CHIMES was better at assessing the superficial (“look”) features of the interfaces, particularly finely detailed inconsistencies. CHIMES performed better in the color and button categories, as is evidenced by the low percentages (<10 and 0, respectively) of items detected only by experts. Higher percentages of items reported only by experts were seen in the other two categories—font and typography. In fact, there were no typography issues identified only by CHIMES. A possible explanation for this result is the fact that typography relates to display logistics. For example, if “FORWARD” was used on one screen, but “fwd” was used on another, CHIMES would detect the case (or font) inconsistency but not the terminology (or typography) inconsistency.

HCI experts were better at assessing logistical (“feel”) features of the interfaces, such as the lack of functional display item grouping. The experts also provided more detailed and specific descriptions of the problems. Inconsistencies/violations detected by at least one expert which were not identified by CHIMES were labeling, digital data format, legend use, and line width. The experts performed better in the typography and color categories; these categories yielded low percentages of items detected by CHIMES only (< 20, respectively). However, there were no button issues (e.g. shadow width) identified by experts only, and the number of issues identified by CHIMES alone was only 5% less than those reported by both. A possible explanation for this result is that CHIMES was better at detecting inconsistencies such as three-dimensional button shadowing—an inconsistency that may be difficult for the human eye to recognize.

3.2.2 Completion Time Results

Completion time for CHIMES was much shorter than the average completion time for the HCI experts. The total completion time for the CHIMES evaluation was 36 minutes. This time takes into account both the time required for the user to set up CHIMES and the time that the application was blocking further input while performing the evaluation. The HCI experts’ completion times ranged from 1 hour, 13 minutes to 2 hours. Their mean time was 1 hour, 33 minutes—approximately 300% of the CHIMES time.

3.3 Discussion

CHIMES cannot replace the evaluation techniques and abilities of an expert, but it appears to be a promising tool to complement HCI expert reviews. A human factors IV&V tool would best be used during the iterative design and testing phase of the design process. Using CHIMES to evaluate objective interface characteristics would reduce evaluation times significantly, since CHIMES took 1/3 of the mean expert time to complete its evaluation. Furthermore, CHIMES has the advantage of detecting problems that are psychophysically difficult for the human eye to detect, such as button shadow consistency. With CHIMES as an evaluation aid, the expert would be spared those time-consuming, monotonous tasks and could concentrate on more cerebral interface issues, such as functionality and interaction styles. Yet another advantage of

incorporating CHIMES into the design process is that it gives the user the ability to modify the interface in real-time.

Before CHIMES is integrated into an IV&V process, we recommended that its capabilities be enhanced to detect the superficial ("look") features that were only reported by the experts (e.g., labeling, legend use, digital data format and even grouping).

4.0 PHASE II

The results of Phase I indicated that CHIMES appears to be a promising complement to HCI expert reviews. Given the success of this first testing phase, experts were then asked to critique the CHIMES user interface itself in Phase II.

4.1 Methods

4.1.1 Participants

Three HCI experts from a contracting organization with JSC participated in this phase of the evaluation. Two of the experts also participated in Phase I. All had HCI graduate education and a minimum of two years of experience in usability testing and HCI reviews. Two experts were involved in interface design. All participants had computer experience. Table 2 summarizes the background information of the HCI experts.

4.1.2 Apparatus and Materials

A SUN workstation with SunOS and Motif was used to run CHIMES. The experts were given an instruction sheet as a guide to operate CHIMES (see Appendix E). In addition, they were provided with computer printouts of the primary screens they would encounter during the evaluation (see Appendix F). The experts were also provided with a questionnaire assessing their general satisfaction with the CHIMES interface design. Each item was rated on a five-point rating scale anchored as 1=Strongly Disagree and 5=Strongly Agree (see Appendix G).

4.1.3 Procedure

Due to limited access to SUN workstations, two of the participants evaluated the CHIMES interface simultaneously, and the third performed the evaluation separately. The experts were given a minimal description of how CHIMES works, and were shown the various screens they would encounter. To test the intuitiveness of the interface, the experts were allowed to explore various menu options and different screens without any time constraints. One of the ISS interfaces was provided for the evaluation so that the experts could see examples of the problems and advice that CHIMES would

provide. The experts were then asked to evaluate the interfaces of the four primary CHIMES screens: Evaluation Control, Advice Index, Problems and Advice, and Case Modifier (which deals with typography). Each expert was provided with computer printouts of the various screens and was asked to make any specific comments directly on these pages. At the conclusion of the evaluation, two of the experts filled out the interface design questionnaire.

4.2 Results

Overall, the CHIMES application was rated positively. The engineer who redesigned the ISS interfaces found TAE+ intuitive and easy to work with. The various menu options were grouped accordingly. The design tools were easy to use and to manipulate. However, the abundance of windows frequently became unmanageable. Furthermore, too many options were included on each window, making the selections hard to distinguish from one another. The average time to recreate an interface was three hours. Of the CHIMES screens themselves, the menu options were found to be grouped appropriately, and evaluators had no trouble locating the appropriate command to perform an action. The layout of the various screens was also rated as acceptable.

Specific CHIMES user interface design issues that the experts identified can be grouped into four categories: Fields/Buttons, Labeling, Feedback, and Procedures.

4.2.1 Fields/Buttons

- 1. Issue:** The scroll bars were too small or inconspicuous. They were not easily seen and were difficult to control with the mouse.

Recommendation: The size of the scroll bars needs to be increased. In addition, scroll bars should be present on each scrollable window, even if the list of items does not exceed the size of the window.

- 2. Issue:** Some icons were difficult to locate. For example, the white hand-shaped pointers, which appear on the display to highlight inconsistencies, are difficult to locate on a "busy" display.

Recommendation: The color of the icon needs to be distinguishable from the background display. One suggestion is to have a blinking icon.

- 3. Issue:** The shapes of buttons used to select similar options were not consistent across screens or even within a screen (e.g., Help, Case Modifier options).

Recommendation: Buttons used to select similar options need to be the same shape and size.

4. **Issue:** Buttons used to select a particular function were not in the same location across screens (e.g., Help appears in the upper right corner on one screen and in the bottom right on a different screen).

Recommendation: Buttons with the same label need to have a fixed location across screens.

4.2.2 Labeling

1. **Issue:** No shortcuts were provided for menu selections.

Recommendation: Since the user will be selecting similar menu options for each display, the labels should also indicate shortcut keys (such as "e" for Evaluate) to avoid using the mouse for function selection.

2. **Issue:** The various problems indicated in the Advice Index window need to be distinguished from one other.

Recommendation: Include an indication of whether the problems are of equal importance through the assignment of a weight or a scale rating to each problem.

3. **Issue:** A "Close" button did not appear on every screen (e.g. Case Modifier).

Recommendation: The "Close" button should appear on every screen and should remain in the same relative spatial location across screens.

4.2.3 Feedback

1. **Issue:** After commanding CHIMES to evaluate a display, no feedback or status indicator was provided. The user had no way of knowing that CHIMES was indeed evaluating or how long the evaluation would take.

Recommendation: Status indicators that could be used include a watch icon, an hourglass icon, or a "percent completed" indication, either pictorial or numerical.

2. **Issue:** The status indicator for selecting an option is ambiguous. For example, the gray "selected" mode is not clearly distinguishable from the black "unselected" mode.

Recommendation: Checkboxes should be filled in (black) to indicate a selection and blank (white) to indicate that the option is not selected. Checkmarks could also be used to indicate the options chosen.

3. **Issue:** The CHIMES prototype did not "deselect" options that are unavailable, such as the Print and Help capabilities. It was discovered that selecting one of these functions would freeze the application, and the user was forced to quit and re-invoke CHIMES.

Recommendation: Feedback needs to be provided for "unselectable" items (e.g. by graying out unavailable options).

4. **Issue:** The Advice Index screen did not provide feedback on the problems already viewed.

Recommendation: It would be helpful if a checkmark, or other such indicator, were used to display the status of problems the user had already viewed.

4.2.4 Procedures

1. **Issue:** Some functions require a single-click to be activated, while others require a double-click.

Recommendation: The mouse function which activates a button should be consistent, preferably a single-click.

2. **Issue:** Some selection functions are activated by highlighting the item (e.g., advice index), while others require more than one step (e.g. an "OK" button).

Recommendation: An intermediate "OK" step should be a common feature on all of the selections. This would provide a chance for the user to change the option selected, as well as consistency.

3. **Issue:** Some of the icons were ambiguous. For example, it was not clear why the hand-shaped pointers were appearing on the display.

Recommendation: More on-screen descriptions of the actions performed by each selection are needed.

4. **Issue:** It was not always clear where or how to proceed through the advice screens.

Recommendation: It would be beneficial to provide brief information on each screen informing the user how to proceed in order to perform a certain action.

The remainder of the problems identified by the experts were due to the prototype nature of the tool. First, some of the menu options are not yet enabled (e.g. Print, Help). In addition, the user is not able to return to the Advice Index window after selecting a particular problem for additional advice and modification. The user is forced to quit the application and re-invoke CHIMES. Such problems were noted but were not of primary concern for this evaluation.

A summary of the Phase II questionnaire results appears in Figure 3. CHIMES was rated as acceptable in most of the categories. The application ranked highest in intuitiveness of layout and lowest in providing sufficient feedback. In general, evaluators indicated that they would use such a tool given the inclusion of more guidelines and more detailed advice. In some cases, it would be beneficial to also include a reference or suggestion, such as "the minimum button size should be..." informing the user of specific guidelines and standards.

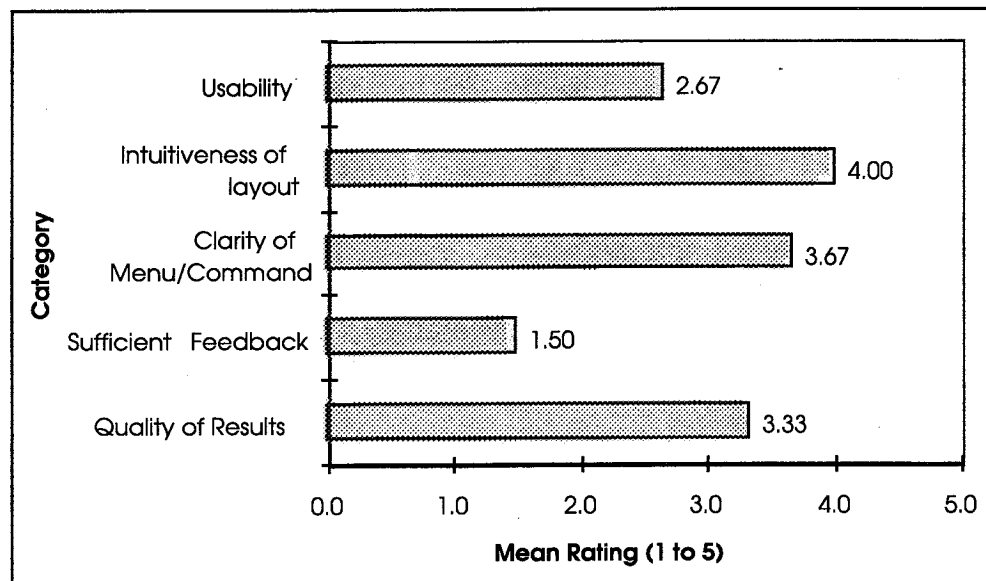


Figure 3: Summary of results from the Phase II questionnaire.

4.3 Discussion

CHIMES is a promising tool to complement HCI experts' reviews and evaluations. Especially if it is recognized for its performance as a "guidelines check," akin to a spell check. Some designers are under the impression that such tools can perform high-level cognitive functions and are disappointed with the concentration on physical features, such as button size and line width. However, research has shown that

designers are often too involved with the overall structure or layout of an interface to notice inconsistencies in these important surface features. In addition, such a tool can identify inconsistencies, such as button shading, that are barely noticeable to the human eye. Correcting such inconsistencies may be critical for complex or crowded displays where there is limited space. Furthermore, it is important that the tool be general, so that it is applicable to interface designs across various projects.

5.0 CONCLUSIONS

Independent verification and validation is one way to ensure quality user interfaces for new software throughout a system. Research indicates that an IV&V tool which assesses the degree to which human-computer interfaces meet human factors standards across each project and program would be extremely beneficial to designers. In addition, the establishment of a human-computer interface IV&V methodology will improve HCI design quality, reduce cost and schedules, and maximize productivity.

Clearly, progress is being made toward automated tools for interface checking, and the utilization of a tool that can accomplish some objective checks seems quite feasible. However, IV&V tools have a long way to go before they are a high-level expert system for human-computer interfaces.

6.0 RECOMMENDATIONS FOR FUTURE WORK

Incorporation of a human factors IV&V methodology into the user interface designs of complex systems, such as ISS core system displays, will ensure the consistency and compliance of all interfaces to NASA's HCI standards. The automated IV&V tool known as CHIMES, reviewed as part of this project, shows promise in terms of evaluating the "look" features of complex user interfaces. However, the following design recommendations need to be implemented in addition to the interface issues listed in Section 4.2:

1. Include more guidelines (add more "look" features, such as labeling and button placement).
2. Give reference to or suggestions based on specific standards.
3. Incorporate criticality ratings for each of the HCI issues identified.
4. Provide "interpreter(s)" to translate any interface to the format that could be read by the tool.

If the CHIMES prototype can be modified by its developers, it is recommended that it should be included in the human factors IV&V process as a complementary tool. A good first-hand application will be verification and validation of both ISS core system and payload displays.

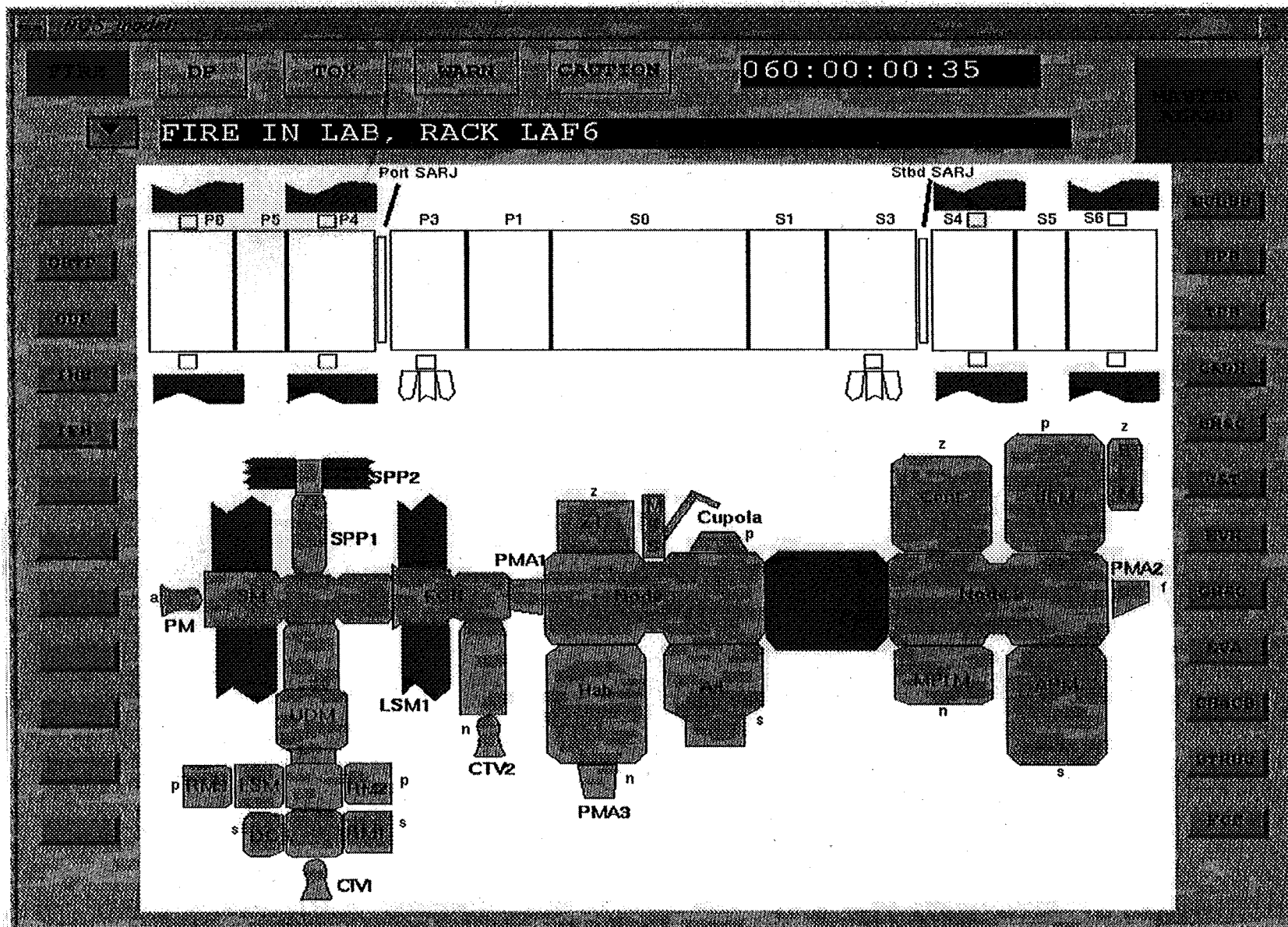
REFERENCES

- Baecker, R. M., Grudin, J., Buxton, W. A. S., & Greenberg, S. (1995). *Human-Computer Interaction: Toward the Year 2000*. San Francisco, CA: Morgan Kaufman Publishers, Inc.
- de Souza, F. & Bevan, N. (1990). The use of guidelines in menu interface design. *Proceedings IFIP INTERACT '90* (Cambridge, UK, 27-31 August), 435-440.
- Harris, E. Personal communication. 1996.
- Jiang, J., Murphy, E. D., & Carter, L. E. (1994) Computer-human interaction models (CHIMES). Technical Paper no. DSTL-94-00. Greenbelt, Maryland: NASA Goddard Space Flight Center.
- Lewis, R. O. (1992). *Independent Verification and Validation: A Life Cycle Engineering Process for Quality Software*. New York: Wiley.
- Lowgren, J. & Lauren, U. (1993). Supporting the use of guidelines and style guides in professional user interface design. *Interacting With Computers*, 5, 385-396.
- Lowgren, J. & Nordquist, T. (1990). A knowledge-based tool for user interface evaluation and its integration in a UIMS. *Human-Computer Interaction--INTERACT '90*, 395-400.
- Lowgren, J. & Nordquist, T. (1992). Knowledge-based evaluation as design support for graphical user interfaces. *Proceedings of CHI '92*, 181-188.
- Mahajan, R. & Shneiderman, B. (1995). A family of user interface consistency checking tools. Technical paper.
- Mosier, J. N. & Smith, S. L. (1986). Application of guidelines for designing user interface software. *Behaviour and Information Technology*, 5, 39-46.
- Sears, A. L. (1994). Automated metrics for user interface design and evaluation. *International Journal of Biomedical Computing*, 34, 149-157.

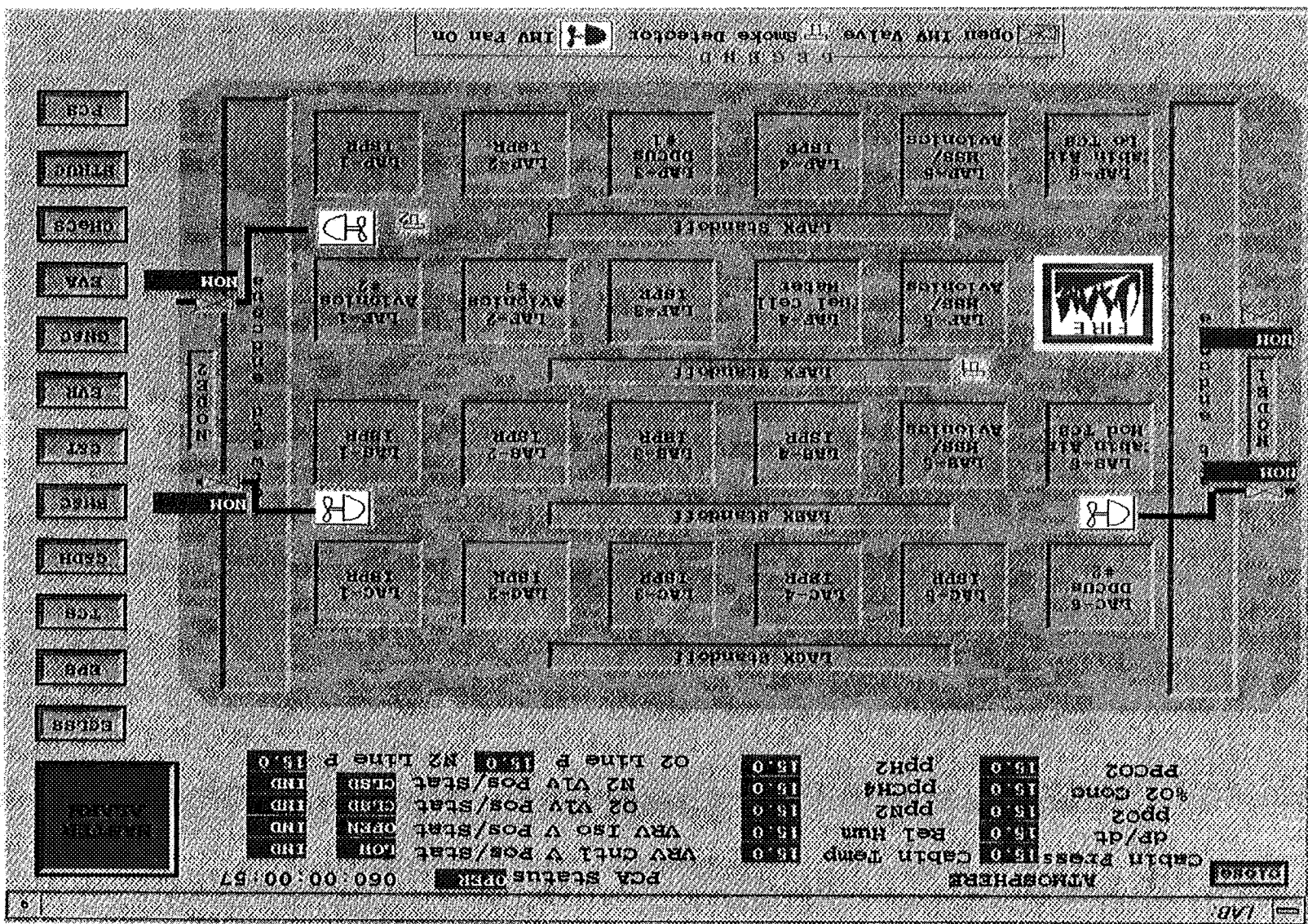
- Shneiderman, B., Chimera, R., Jog, N., Stimart, R., & White, D. (1995). Evaluating spatial and textual styles of displays. Technical paper.
- Siochi, A.C. & Hix, D. (1991). A study of computer-supported user interface evaluation using maximal repeating pattern analysis. *Proceedings of CHI '91*, 301-305.
- Tetzlaff, L. & Schwartz, D. R. (1991). The use of guidelines in interface design. *Proceedings of CHI '91*, 329-333.
- Thovtrup, H. & Nielsen, J. (1991). Assessing the usability of a user interface standard. *Proceedings of CHI '91*, 335-341.
- Tullis, T. S. (1983). The formatting of alphanumeric displays: A review and analysis. *Human Factors*, 25, 657-682.
- Whitmore, M. & Berman, A. H. (1996). *Common Ground for Critical Shuttle and Space Station User Interfaces: An Independent Verification and Validation Approach*. Poster presented at the Computer Human Interface '96 Conference (CHI '96).

Appendix A

Phase I: ISS Interface Designs



Top-Level PCS Display



[illegible]

Line	Description	Amount	Balance
100.00	CHECKNOIT		
\$8 5554 \$8	ht node	\$1	
\$0	payload id sequence stop gating	\$0	
P1			
P3			
P4P5 P6			

ON AIRLINE

MODE III

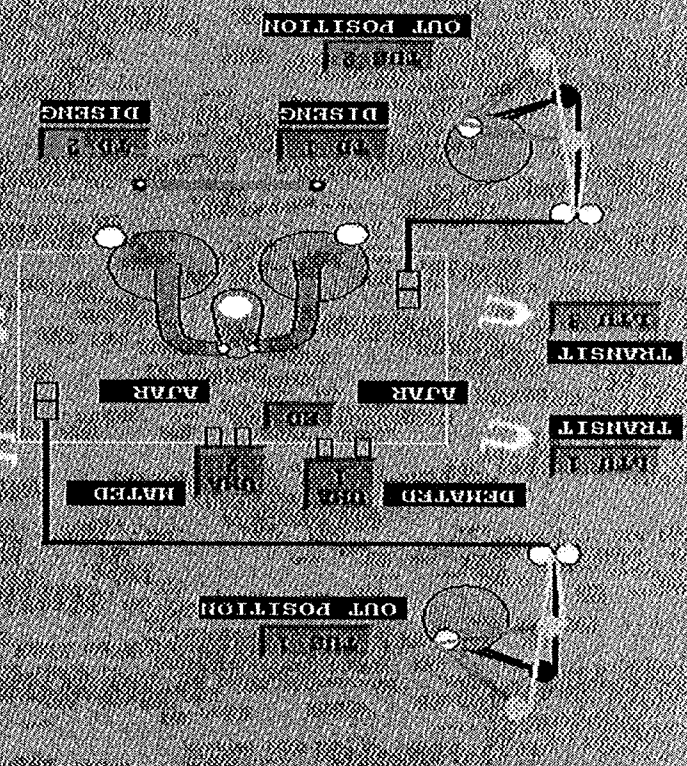
RECORDING	PAYROLL ID		
	CHECK DATE	CHECK OUT	CHECK IN

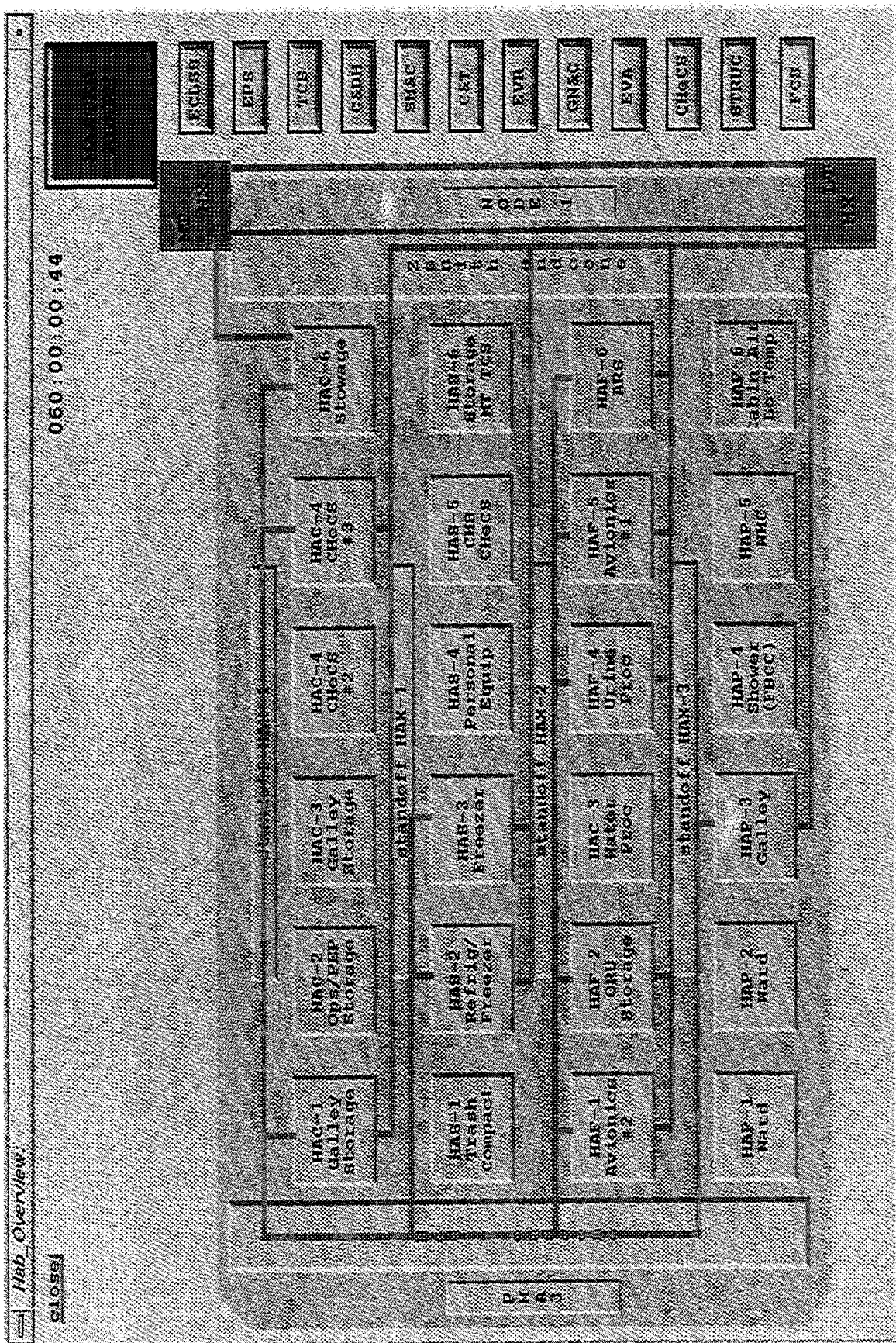
IN CLARENCE

REAR	ENGAGED	DISARMED	ENGAGEMENT DRIVE
REAR	UNLATCHED	LATCHED	LATCH CONTROL

SWITCH PWR STRING	PRIME	BBB	BBB
ABSOLUTE POSITION		0	0
RELATIVE POSITION		0	0

CONFIGURE HT		
GRU ID	0	GRND
MIN HEATER TEMP	0	GRND
MAX HEATER TEMP	0	GRND
MAX VELOCITY	0	GRND
MAX ACCELERATION	0	GRND
CURRENT LOCATION	0	GRND





Habitation Module Overview Display

Appendix B

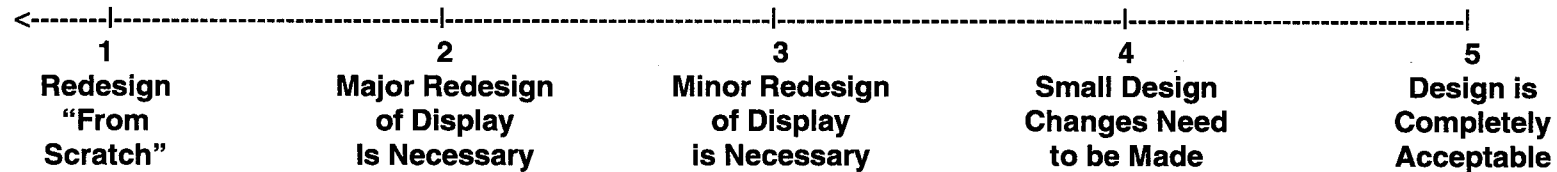
Phase I Comment Table

As you check each screen for compliance, there are certain general dimensions you should keep in mind: readability, clarity, consistency, helpfulness, and ease of comprehension. Give design recommendations, if you have any.

NOTE 1: If you brought reference materials with you for these evaluations, note when you used them by placing an asterisk (*) next to your comment for that interface characteristic.

NOTE 2: If you have any other HCI-related comments about these displays as you're working, note them under "Other Comments."

NOTE 3: As you comment on the compliance of each interface characteristic, rate the display only relative to that characteristic using the following scale and place your rating in the appropriate box. If you give a low rating (< 3), explain your rating.



SCREEN #	INTERFACE CHARACTERISTIC			
	Fonts	Lines	Colors	Buttons
1				
	Other Comments:			

Appendix C

Phase I Questionnaire

IV&V Evaluation Questionnaire

PART 1: Past Experience

1. How many different types of computer systems have you worked on? Respond by placing the number of years of experience with each system in the blank provided. Place an asterisk next to the system on which you prefer working.

_____ Mac	_____ UNIX
_____ Microsoft Windows	_____ IBM/DOS
_____ Microsoft Windows 95	_____ Other: _____

2. What is your level of experience in usability activities? Respond by placing the number of times or years you've been involved in each type of activity in the blank provided.

_____ Participated as a subject in software usability testing	_____ Checked software interfaces for compliance to HCI standards
_____ Conducted software usability testing	_____ Performed display reviews
_____ Designed software usability experiments	_____ Performed rapid prototyping
_____ Designed software interfaces	

3. How familiar are you with the space station displays including both the SSF and ISS? Respond by putting a check mark on the relevant topics.

_____ Conducted evaluation of ALL subsystems	_____ Conducted evaluations of one subsystem: _____
_____ Attended crew reviews	_____ Participated in displays & controls team meetings
_____ Attended display demonstrations	_____ Other: _____

4. How many years of HCI related background do you have?

_____ Education in HCI	_____ Education in human factors
_____ Technical training in HCI	_____ Work experience
_____ Other: _____	

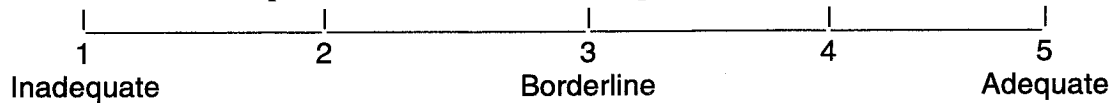
5. How much familiarity do you have with the following documentation? Respond by marking the appropriate selection.

	Familiar with:	Extensive use:
Macintosh Human Interface Guidelines	_____	_____
Object-oriented Interface Design	_____	_____
Open Look; Graphical User Interface Application Style Guidelines	_____	_____
Other: _____	_____	_____

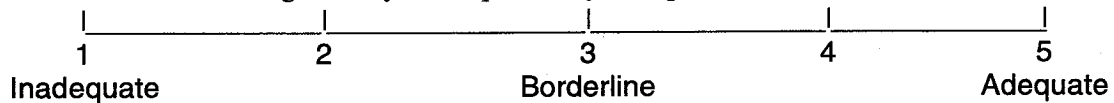
PART 2: Overall Reaction

Use the scales listed below each question to answer the following questions. Circle the number that best matches your opinion. Place additional comments below the question in the space provided.

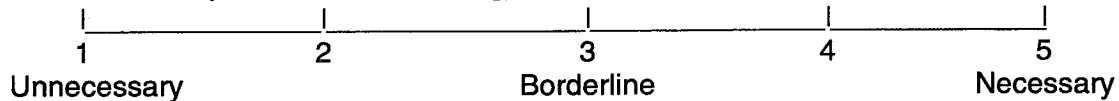
1. Was the total time provided for the evaluation adequate for you to work at your own pace?



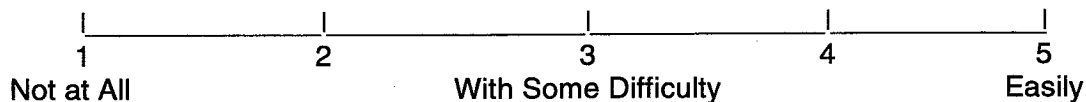
2. Was the instruction given to you adequate for you to perform this evaluation to the best of your ability?



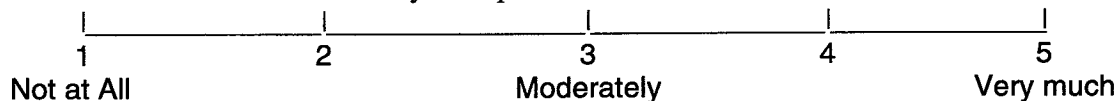
3. How necessary were the references to perform this evaluation to the best of your ability?



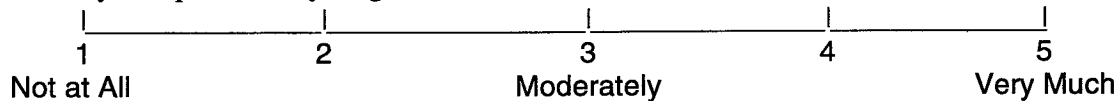
4. How well were you able to maintain the same level of detail in your display analyses throughout the evaluation?



5. How much did this task utilize your expertise in the field of HCI?



6. Did you experience any fatigue or discomfort after the evaluation?



Appendix D

Results of Experts' and CHIMES' Analyses of the ISS Displays

Results of HCI Experts' and CHIMES' Analyses of the ISS Displays

The results of the HCI experts' analyses are presented in Sections 1.0 through 5.0. Those items which were identified by CHIMES only are presented in Section 6.0. Each display was critiqued in terms of its "look" features (i.e., physical layout) as well as the "feel" features (i.e., interface logistics), which were detected by the experts. In addition, they were asked to rate the criticality of display issues on a scale of 1=Redesign from Scratch to 5=Design is Completely Acceptable. It is important to note that the experts were not asked to identify "feel" features nor were they asked to make design recommendations. However, these items were recorded by the experts as part of their analyses, and the complete results are presented below.

1.0 The Top-Level PCS Display

This display shows a schematic of all ISS modules, with a visual warning of a fire occurring in the U.S. Lab module.

1.1 Physical Layout

Fonts were found to be inconsistent, with proportional and non-proportional fonts mixed together, and no indication of any apparent grouping. The font appeared grainy, and the use of several font sizes and styles was not necessary.

Color coding, if any existed, was unclear. For example, two different shades of red were used. One shade was used as a warning of a fire in the U.S. Lab and the other as a standard display color. In addition, the significance of dark blue was unclear. Multicolored text was considered unnecessary, and white was thought to be a poor choice for the display background color.

Regarding the use of buttons, it was noted that inactive buttons should be grayed-out and that unlabeled buttons had no clear purpose. The organization of buttons into columns on the right- and left-hand sides of the screens was a non-optimal choice; instead, buttons should be grouped by function. In addition, it was unclear what clicking on the "MASTER ALARM" button would do. It was also noted that the clock at the top of the display needed a label.

With respect to line usage, the meanings of lines of various thicknesses were unclear. On the ISS schematic itself, the modules were not spaced consistently, creating some thicker lines between modules that probably were insignificant. In addition, the visual fire warning should have a black border. Lastly, the lines that were labeled "Port SARJ" and "Stbd SARJ" could have been considered part of the physical structure or could have been acting as arrows. Since they were probably arrows, actual arrowheads would have made this point clearer.

1.2 Interface Logistics

It was assumed that the upside-down triangle brought up a pull-down menu when the mouse button was depressed. If this was so, and the information that popped up was relevant to the fire, that information instead should have been automatically presented on the display with no user action required.

An "X" or a fire icon should mark the specific location of the fire within the module. Graphic elements should be improved to better represent the truss structure and module configuration. "CAUTION" and "WARNING" buttons should be set apart from the emergency buttons (such as atmosphere depressurization - DP and atmospheric toxicity - TOX). Furthermore, if the two columns of buttons are not associated with the row of buttons at the top of the display, there needs to be a horizontal line separating the row from the columns.

Generally speaking, it is not clear what action needs to be taken to handle the fire in the Lab, nor is it clear which button to click on to be provided with information on the necessary corrective action. Ratings for this screen ranged from 3 to 3.5.

2.0 The Lab Fire Display

In this display, each rack in the Lab module is represented by a button, and a fire icon covers the system that is on fire. Atmosphere and valve information is provided at the top of the screen.

2.1 Physical Layout

It was again suggested to use a less grainy font. Although some button titles did not fit on the buttons themselves and were cut off on the button edges, use of a larger font was suggested. The larger point size could be compensated by using a font with characters more closely spaced together. The use of two font sizes was unnecessary, however the consistent use of one font was good. The smoke detector numbers were illegible, and vertical labels should have been avoided if possible. The information field labels at the top of the display should be left-justified rather than centered or right-justified, and units for that information should be provided. The fire icon covers up the name of the rack which was on fire, and the user was required to search adjacent racks to determine which one is on fire. The top of the display was very cluttered; grouping delimiters would help.

Since the fire icon was both red and yellow, it was unclear whether it was a caution or an emergency warning. If the smoke alarms have been activated by the fire, this should be shown with some sort of color coding. Color shading was considered good, however the choice of blue for some buttons was unnecessary. Three-dimensional buttons with an identical background color were completely legible. It was noted that they would not have interfered with the user's attention when the color of a display item changed to indicate a new condition. Text was legible on all background colors. There seemed to be no reason for multiple icon colors.

Some buttons had borders and some did not. Although it was reasonably clear which display items were buttons, their functions were not clear. Again, the clock should be labeled. The column of buttons on the right side of the display should be separated from the module. The "close" button in the upper left-hand corner is unnecessary. The endcone buttons do not represent endcones well, and the node buttons should be placed outside the outline of the Lab in the schematic.

The thick black lines were seen as pipes by two of the four experts. These lines obscured parts of some button labels, and it was suggested that the display items be repositioned so that this does not occur.

2.2 Interface Logistics

The legend should show what the valve and the fan look like when off as well as when on in order to match the icons in the schematic with those in the legend. It is also unclear if it is significant that one fan is facing the opposite direction.

Again, the actions required for handling the fire were unclear, and it was unknown how to find that information in the display. To reduce screen clutter, only the critical atmospheric and valve information should be provided at the top of the display. This display received a consistent rating of 3.

3.0 The Mobile Transporter Home Display

The mobile transporter (MT) will travel on the exterior of the ISS and perform extravehicular orbital replacement unit (ORU) activities. This display provides positional and systemic information on the MT and its activities.

3.1 Physical Layout

With all the available space on this display, a much larger (and consistent) font point size could have been used. The use of a non-proportional font is good, but a less grainy font would have been a better choice. Labels with all upper-case letters should have been reserved for those items for which users would initially scan the display. Furthermore, the labels along the top of the display do not align with anything.

Too many colors were used. Yellow should be reserved for warning, and it was unclear why some items were sky blue or blue-green. Colors should have helped group objects or separate areas of the screen from each other. To truly emphasize the "E-STOP" button, both the button and the label should be of different colors than the display background. The MT structural graphics were difficult to see, since they were light and dark gray on a gray background. The graphics should have some outline shading to separate them from the command and output fields.

Some of the buttons were too small and got lost in the display. They were not well-grouped and did not stand out well against the background color, although the three-dimensional effect allowed them to be somewhat easily identified as buttons. The "off"

button should not have been isolated in the upper left-hand corner; if it was equivalent to the "close" button in the previously discussed display, then there should have been consistent labeling. The "LATCHED"/"UNLATCHED" and "DISENGAGE"/"ENGAGED" buttons should be shifted to the left to put some space between them and the "SEND" button. The way it was organized, the "SEND" button appeared to be a third option. The "E-STOP" label was found to be too terse; whatever "E" stands for (probably "emergency"), there was enough real estate on that button for it to be spelled out.

Some successful usage of lines to separate functional button groups is seen but not enough; boxes or borders would do a better job of spatial separation. On the truss diagram at the top of the display, some lines are thicker than others; this appears to be an alignment accident and should be corrected.

3.2 Interface Logistics

The mixed use of action and state terminology in the "DISENGAGE"/"ENGAGED" button pair is confusing. "DISENGAGE" suggests that clicking on the button causes the MT to become disengaged (action terminology). "ENGAGED" also suggests that clicking on the button causes the MT to become disengaged (state terminology). Since the "LATCHED"/"UNLATCHED" button pair uses state terminology, it is suggested that "DISENGAGE" be changed to "DISENGAGED." There also appeared to be no clear feedback in the graphics regarding the MT's latched and/or engaged status.

Status buttons and displays should be separated from command buttons and displays. The multiple "SEND" buttons seemed unnecessary; one "SEND" button for the entire display was considered sufficient. One expert also brought up the point that after selecting a command (such as "ENGAGED" or "LATCHED"), the user should not be required to send that command in a second step using the "SEND" button. Instead, once a command is selected, it should automatically be sent to the MT.

Ratings for this display ranged from 1 to 2.

4.0 The Habitation Module Overview Display

This display resembles the Lab Fire display, with each rack in the Habitation (Hab) module represented by a button. No fire is shown on this display, and there are no information fields at the top.

4.1 Physical Layout

The font was considered too small and too grainy, with the characters spaced too far apart. Some button labels did not fit on the buttons and were cut off. A font with less serifs would have been a better choice.

The meaning of dark blue buttons and areas in the display was unclear. One expert thought it might represent either a fluid or nonselectable display items. In addition,

black text was difficult to read on the dark blue background. Color coding could have been used to group items more logically.

Almost every item on the screen looked like a button, but it was unclear what purpose each served. Only the buttons in the right-hand column were outlined in black, and the experts wondered if that held some meaning. The "MASTER ALARM" button should have been more isolated in the display, by using borders or by physically relocating the button. The Node 1 and PMA3 buttons are located inside the outline of the Hab module when in reality, they exist outside of the Hab. The clock in the upper right-hand corner is just floating; it needs to at least be labeled.

The lines on the schematic should be labeled, because it is unclear what they represent. A legend would help, especially if each line thickness represents something. Also, the lines should not be obscuring parts of button labels.

4.2 Interface logistics

The "MASTER ALARM" button caused some confusion. It was difficult to tell if it had been activated, needed to be activated by the user, or simply sat in the corner of the screen for when an emergency might occur.

If this display, or the buttons within it, are supposed to provide the user with information, that point needs to be made more clear. There is too much text to force the user to read it all before choosing an action.

Furthermore, if the buttons in the right-hand column take the user to other system displays, that should be stated in some manner. The relationship between those buttons and the Hab module should be clarified.

Improving item grouping and color coding would make this display much more usable and would help to convey information better. This display received ratings ranging from 3 to 4.

5.0 Consistency Across Screens

The same font was used consistently, but size and case were highly inconsistent within and across displays. The use of lines to structure displays and to box similar buttons was consistently absent. Line thickness differed, and it was unclear whether this served some purpose. While some buttons had black borders, others did not.

The gray background color and module color shading were consistent, but all other uses of color were not. Information fields and button text were either black or dark blue. White lines were used to represent different items on two of the displays. The "MASTER ALARM" button colors changed on the displays; it was unclear whether this was a consistency problem or whether the color signified alarm status.

Buttons were fairly consistent but sometimes were two-dimensional and sometimes three-dimensional. Only one display had a legend, but all displays could have benefited from one. Too many sizes of buttons were used. Labeling was inconsistent. For example, on one display "FWD" was used and on another, "forward" was used. Lastly, the number of decimal points in information fields was inconsistent.

One expert commented that there was no general, common "feel" to all the displays. This expert believed, therefore, that learning and understanding one display would not help the user to become familiar with all the other displays.

6.0 CHIMES Results

This evaluation was repeated, using the CHIMES application in place of an HCI expert. Some interface design issues were detected by both, and some issues were only identified by CHIMES or by the experts. To avoid redundancy, this section will only discuss those issues detected by CHIMES alone. Sometimes, CHIMES detected issues in one display that were identical to those identified by the experts in other displays (Whitmore & Berman, 1996).

When CHIMES points out an interface problem, it is always accompanied by an explanation of the problem in human factors terminology. On the Lab Fire display, CHIMES noted that more than three fonts were used and pointed out that fonts should be varied only when it is meaningful—not for purely decorative purposes. Also on this display, CHIMES recommended a different foreground color. The rationale was that black, blue, or white would be the most legible, high-contrast colors on a medium-gray background. The experts missed this point on the Lab display but noted it on some of the others.

On both the Lab Fire and the Top-Level PCS displays, button width, height, and shadow inconsistencies were observed by CHIMES. These attributes should remain constant, unless different functions are assigned to buttons of different size and shadowing. Button shadow inconsistency was also detected across all displays.

7.0 Discussion and Conclusions

The mean rating that the HCI experts assigned to each display ranged widely between 1 and 5. The MT display received the lowest rating of 1.5, and the Hab Module display received the highest: 3.5. The reasons behind the ratings are made apparent in Table D-1, which outlines the major problems encountered by both the experts and by CHIMES in each display.

Table D-1: Major Problems Encountered With Each Display

Display	Problem Description	Mean Rating
Top-Level PCS	<ul style="list-style-type: none">• Inconsistent font and font size• Ambiguous color coding• Unclear button functions and organization• Inconsistent line thickness on graphics• Unclear pull-down menu icon• Location of fire too general• No information provided regarding actions required to handle fire (or how to find same information)	3.25
Lab Fire	<ul style="list-style-type: none">• Inconsistent font size; average font size too large• Ambiguous color coding; too many icon colors• Unclear button functions and organization• Some lines obscured portions of button labels• Legend not descriptive enough—needs more objects• No information provided regarding actions required to handle fire (or how to find same information)	3.00
Mobile Transporter	<ul style="list-style-type: none">• Inconsistent font size; average font size too small• Different label cases not used appropriately• Some labels not aligned with related objects• Too many colors• Color of some graphics too close to background shade• Unclear button (and display graphic) organization• Inconsistent line thickness on graphics• Mixed use of action and state terminology on some button labels• Status and command buttons not separated	1.50
Habitation Module	<ul style="list-style-type: none">• Font size too small• Ambiguous color coding• Unclear button functions and organization• Inconsistent line thickness• Some lines obscured portions of button labels• More instructional text required	3.50

Some critical inconsistencies were identified during these evaluations. With respect to color coding, red and yellow have widely accepted meanings (emergency and caution, respectively), and those meanings need to be respected. The introduction of blue in some displays had no apparent meaning, since there is no widely accepted coding for blue. As CHIMES stated, if colors do not convey meaning or functional relationship, they should not be used solely for decoration.

Functional grouping of buttons was suggested to reduce screen clutter and to make it easier for the user to scan displays for required buttons. More explanatory cues regarding the functions of different buttons should be introduced into the displays; on some displays, every item appeared to be a button, and this layout may overwhelm the user.

IV&V of NASA's critical and high-risk ISS core system displays is a profound task. It involves the analysis of both mission control and vehicle interfaces from the ISS program. The program consists of ten main display development teams (e.g., Systems, Operations, Safety & Mission Assurance), each of which is subdivided into at least five smaller teams. These teams are all responsible for different systems that are developed in parallel. Inter-team communication regarding standardization of software user interfaces is quite a challenge. The IV&V process is one possible mechanism for achieving common ground between these diverse systems.

The goal of conformance of space-related critical software displays to human factors and HCI standards is an important one. Such a common ground will indeed provide a unifying framework for future crews of the ISS. Standardizing interfaces will allow the crews to perform necessary tasks without having to shift mental paradigms between systems.

Appendix E

Phase II Testing Instructions

CHIMES Usability Testing

Instructions for the Experimenter

STARTUP

Turn on the monitor.

Put the cursor on an empty space and hold down the right button.

Click "Open" to open a window for the "mysnap" program.

Downsize that window.

Go to the other window that was already open and type "chimes."

SHUT DOWN

When you are ready to leave, click in an empty space on the screen with the right mouse button and choose "exit" to log off.

Turn off the monitor only.

Instructions for the Evaluator

Click "close" when prompted in the CHIMES - Acknowledgments Window

DISPLAY (FILE) SELECTION

Click on "file" and "open" and "resource file" in the CHIMES - Evaluation Control window.

In the CHIMES - Resource Files window, select resource file "MT.res" from the window on the right.

SETUP OPTIONS

Click "options," "select category" in the CHIMES - Evaluation Control window.

Deselect "OSF/Motif", and hit "OK."

RUN CHIMES

Click "evaluation" in the CHIMES - Evaluation Control window.

Record # of advice items.

Request advice on the "Typographic inconsistency" in the CHIMES - Advice Index window.

Select "modify."

QUIT CHIMES

When you are finished, click "file" and "exit" in the CHIMES - Evaluation Control window.

Critique the following windows:

CHIMES - Evaluation Control

CHIMES - Advice Index

CHIMES - Advice Problems and Advice

CHIMES - Case Modifier

Fill out the questionnaire.

Appendix F

Phase II CHIMES Screens

CHIMES - Advice Index

Please select an item in the index to view the advice, then you can either delete the advice, select another index item or close this panel.

Advice Index:

Push button typographic inconsistency

Push button width inconsistency with a panel

Background inconsistency between panel and items

Delete

Close

Help

CHIMES - Problems and Advice

Please review the problem and the advice. Then you can modify the design if the 'Motif' button is not dimmed. An empty problem statement means the advice is a tip.

Problem

The shadows of push buttons in this panel are not consistent.


Advice

Push button shadows should be the same within a panel unless different shadows convey functional relationships between the push buttons.


Modify

Close

Help


CHIMES - Case Modifier

Please select scope and typography to make changes.



Scope Control:

☐ These Items
☐ This Panel
☐ All Panel


Typography Selections

☐ ALL CAPITALS

☐ Mixed Cases

☐ all lower cases

OK


CHIMES - Evaluation Control

File Options Customizations Evaluation Help

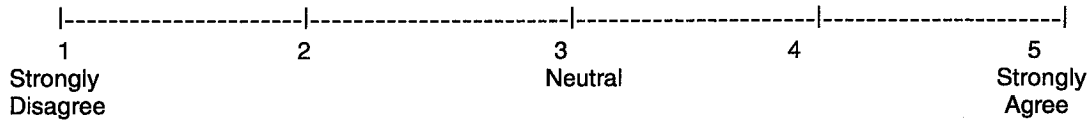
Resource File: LAB.res
Guidelines File: rules.clp

Appendix G

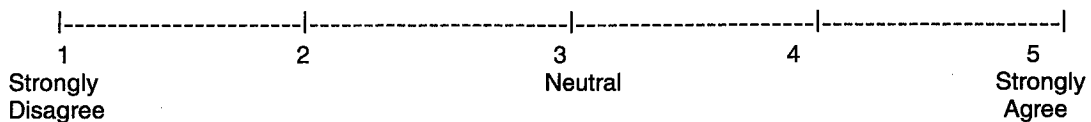
Phase II Questionnaire

POST-EVALUATION QUESTIONNAIRE:

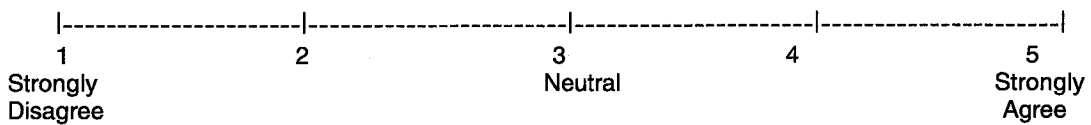
I found it easy to learn how to use the application.



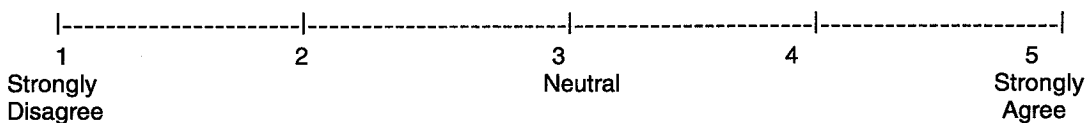
The menu headings and command labels were easy to understand.



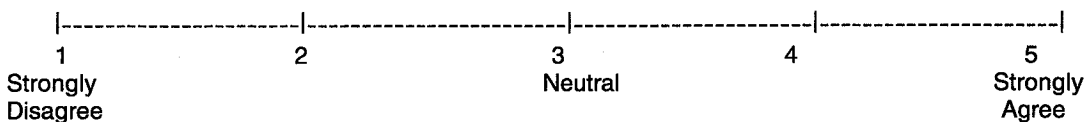
The menus and menu options were grouped appropriately.



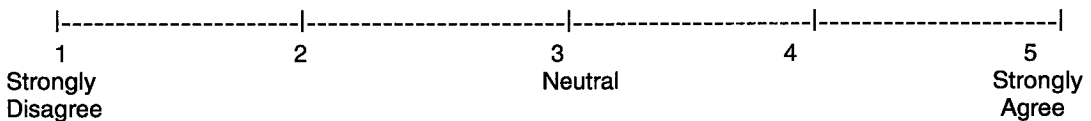
I did not have trouble finding the appropriate command to perform an action.



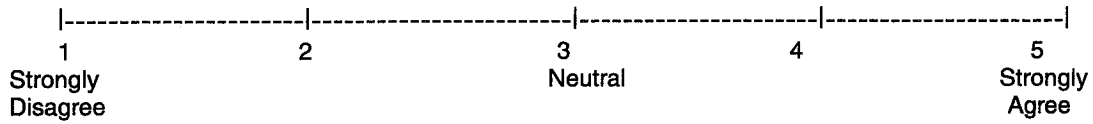
I was satisfied with the amount of feedback I received during the analysis.



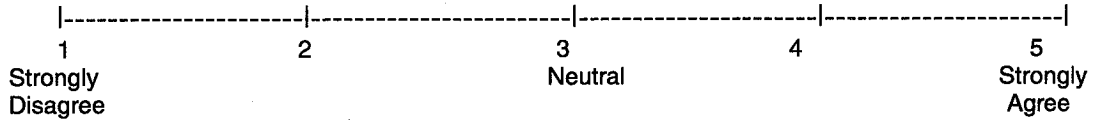
I was satisfied with the amount of information the application provided about each problem.



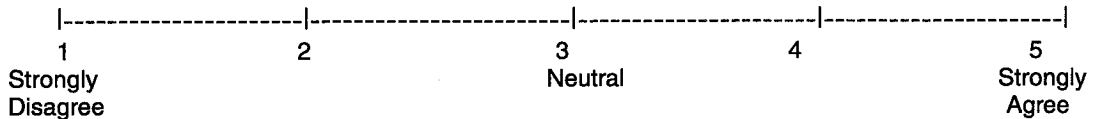
The layout of screens was helpful in comprehending the information presented.



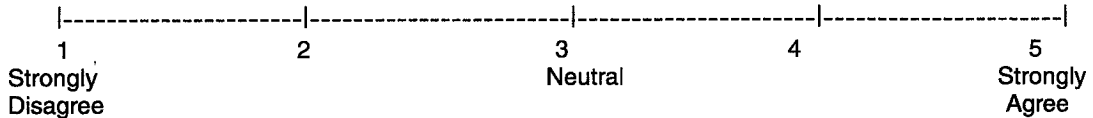
I was satisfied with the results of the analysis.



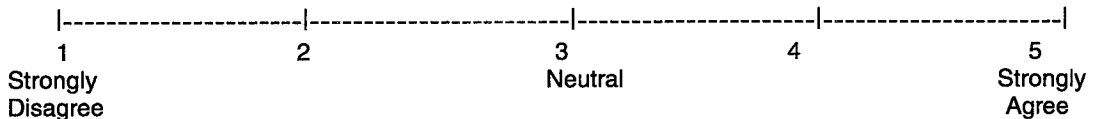
I am confident that the application found most of the interface problems.



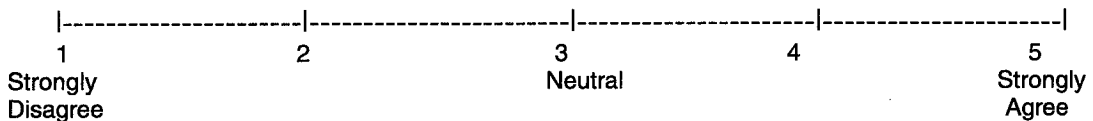
The application identified problems that I found to be extraneous.



I think that there are too many options and special cases for this application to be helpful.



I would use this application in the future.



If not, what improvements would have to be made?

REPORT DOCUMENTATION PAGE			Form Approved OMB No. 0704-0188	
Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302, and to the Office of Management and Budget, Paperwork Reduction Project (0704-0188), Washington, DC 20503.				
1. AGENCY USE ONLY (Leave Blank)	2. REPORT DATE December 1996	3. REPORT TYPE AND DATES COVERED NASA Technical Paper		
4. TITLE AND SUBTITLE Independent Verification and Validation of Complex User Interfaces: A Human Factors Approach		5. FUNDING NUMBERS		
6. AUTHOR(S) Mihriban Whitmore*; Andrea Berman*; Cynthia Chmielewski*				
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) Lyndon B. Johnson Space Center Flight Crew Support Division Houston, Texas 77058		8. PERFORMING ORGANIZATION REPORT NUMBERS S-823		
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES) National Aeronautics and Space Administration Washington, D. C. 20546-0001		10. SPONSORING/MONITORING AGENCY REPORT NUMBER TP-3665		
11. SUPPLEMENTARY NOTES				
12a. DISTRIBUTION/AVAILABILITY STATEMENT Unclassified/Unlimited Available from the NASA Center for AersoSpace Information 800 Elkridge Landing Road Linthicum Heights, MD 21090 (301) 621-0390 Subject Category 61			12b. DISTRIBUTION CODE	
13. ABSTRACT (Maximum 200 words) The Usability Testing and Analysis Facility (UTAF) at the NASA Johnson Space Center has identified and evaluated a potential automated software interface inspection tool capable of assessing the degree to which space-related critical and high-risk software system user interfaces meet objective human factors standards across each NASA program and project. Testing consisted of two distinct phases. Phase 1 compared analysis times and similarity of results for the automated tool and for human-computer interface (HCI) experts. In Phase II, HCI experts critiqued the prototype tool's user interface. Based on this evaluation, it appears that a more fully developed version of the tool will be a promising complement to a human factors-oriented Independent Verification and Validation process.				
14. SUBJECT TERMS human factors engineering, man-computer interface, software engineering, computer systems programs, software tools, program verification (computers)			15. NUMBER OF PAGES 52	
			16. PRICE CODE	
17. SECURITY CLASSIFICATION OF REPORT Unclassified	18. SECURITY CLASSIFICATION OF THIS PAGE Unclassified	19. SECURITY CLASSIFICATION OF ABSTRACT Unclassified	20. LIMITATION OF ABSTRACT Unlimited	